

# *Representativity error for temperature and humidity using the Met Office high-resolution model*

Article

Published Version

Creative Commons: Attribution 3.0 (CC-BY)

Waller, J. A., Dance, S. L., Lawless, A. S., Nichols, N. K. and Eyre, J. R. (2014) Representativity error for temperature and humidity using the Met Office high-resolution model. Quarterly Journal of the Royal Meteorological Society, 140 (681). pp. 1189-1197. ISSN 1477-870X doi: <https://doi.org/10.1002/qj.2207> Available at <https://centaur.reading.ac.uk/34166/>

It is advisable to refer to the publisher's version if you intend to cite from the work. See [Guidance on citing](#).

Published version at: <http://dx.doi.org/10.1002/qj.2207>

To link to this article DOI: <http://dx.doi.org/10.1002/qj.2207>

Publisher: Royal Meteorological Society

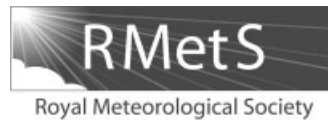
All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the [End User Agreement](#).

[www.reading.ac.uk/centaur](http://www.reading.ac.uk/centaur)

**CentAUR**

Central Archive at the University of Reading

Reading's research outputs online



# Representativity error for temperature and humidity using the Met Office high-resolution model<sup>†</sup>

J. A. Waller,<sup>a\*</sup> S. L. Dance,<sup>a</sup> A. S. Lawless,<sup>a</sup> N. K. Nichols<sup>a</sup> and J. R. Eyre<sup>b</sup>

<sup>a</sup>*School of Mathematical and Physical Sciences, University of Reading, UK*

<sup>b</sup>*Met Office, Exeter, UK*

\*Correspondence to: J. A. Waller, School of Mathematical and Physical Sciences, University of Reading, Berkshire RG6 6AX, UK.

E-mail: j.a.waller@reading.ac.uk

This article is published with the permission of the Controller of HMSO and the Queen's Printer for Scotland.

The observation-error covariance matrix used in data assimilation contains contributions from instrument errors, representativity errors and errors introduced by the approximated observation operator. Forward model errors arise when the observation operator does not correctly model the observations or when observations can resolve spatial scales that the model cannot. Previous work to estimate the observation-error covariance matrix for particular observing instruments has shown that it contains significant correlations. In particular, correlations for humidity data are more significant than those for temperature. However it is not known what proportion of these correlations can be attributed to the representativity errors. In this article we apply an existing method for calculating representativity error, previously applied to an idealised system, to NWP data. We calculate horizontal errors of representativity for temperature and humidity using data from the Met Office high-resolution UK variable resolution model. Our results show that errors of representativity are correlated and more significant for specific humidity than temperature. We also find that representativity error varies with height. This suggests that the assimilation scheme may be improved if these errors are explicitly included in a data assimilation scheme.

**Key Words:** data assimilation; correlated observation error; forward model error

Received 12 September 2012; Revised 25 April 2013; Accepted 27 May 2013; Published online in Wiley Online Library

## 1. Introduction

In data assimilation, model states are combined with observations, making use of their associated error statistics. These are included in the assimilation scheme in the background- and observation-error covariance matrices. The observation-error covariance matrix can be split into three components. One contains information on the instrument error, one describes the error in the observation operator, and the third contains information on the representativity error (RE), also known as representativeness or representivity error. We follow Daley (1991, 1993) and define the RE as the error that arises when the observations resolve spatial scales that the model cannot. The RE and the error in the observation operator can be combined into a single error known as the forward model error (Harris and Kelly, 2001; Sherlock *et al.*, 2003) or forward interpolation error (Lorenc, 1986). In more

recent literature (Cohn, 1997; Liu and Rabier, 2002; Janjic and Cohn, 2006), the term RE has been used to describe the forward model error. However we use the definition given by Daley (1993) as in this paper we focus on calculating the error caused by the misrepresentation of small scales. The instrument error is determined for specific instruments under a set of test conditions by the instrument manufacturer or from in-orbit calibration data. Pre-processing the data may also introduce errors, which may often be greater than the instrument noise. These errors may be included either in the instrument error or in the observation operator error, where they contribute to the forward model error.

Previous work has shown that observation error statistics are correlated for certain observation types (Bormann *et al.*, 2002, 2010; Stewart *et al.*, 2009, 2012; Bormann and Bauer, 2010) and it has been suggested that part of the correlation comes from RE rather than the instrument error or errors in the observation operator (Stewart, 2010; Weston, 2011). Until recently it has been assumed that it is too expensive to include correlated observation error matrices in assimilation schemes and that it is only feasible to use a diagonal observation error covariance matrix.

<sup>†</sup>The copyright line for this article was changed on 28 February 2014 after original online publication.

The effect of correlated error is reduced by using techniques such as observation thinning (Lahoz *et al.*, 2010) or superobbing (Daley, 1991), and variance inflation (Whitaker *et al.*, 2008; Hilton *et al.*, 2009). Calculation are also simplified by assuming that the observations errors are the same at each model level (Dee and Da Silva, 1999). Efforts are being made to find methods of reducing the cost of using correlated observation error matrices (Healy and White, 2005; Fisher, 2005; Stewart, 2010; Stewart *et al.*, 2013). Once these methods are in place it will be important to have accurate estimates of the covariance matrices, as these are required to obtain the optimal estimate from any data assimilation system (Houtekamer and Mitchell, 2005; Stewart *et al.*, 2008). It is therefore important to understand how to estimate RE.

Despite the difficulties in calculating correlated error, there have been some attempts. The Hollingsworth and Lönnerberg (1986) method has been used to calculate the statistics of the innovations. A method proposed by Desroziers *et al.* (2005) makes use of information from the first guess and analysis departures and yields an approximation to the observation error covariance matrix. Once the innovation statistics or the observation error covariances have been calculated, the background and/or instrument error terms can be subtracted to leave an approximation of forward model error for specific observing instruments. Other methods (Daley, 1993; Liu and Rabier, 2002) assume that observations can be written as a projection of a high resolution model state on to observation space with the RE being the difference between this high-resolution projection and the model representation of the observation. Many of these approaches yield a static approximation of RE, but Janjic and Cohn (2006) show in theory that it is state dependent and correlated in time.

Work has been carried out (Stewart *et al.*, 2009, 2012; Stewart, 2010; Bormann and Bauer, 2010) to calculate estimates of the full observation error covariance matrix. They show that the observation-error covariance matrices for observing instruments such as IASI (Infrared Atmospheric Sounding Instrument), AMSU-A (Advanced Microwave Sounding Unit-A), HIRS (High-Resolution Infrared Sounder) and MHS (Microwave Humidity Sounder) contain significant correlations. In particular, the correlations for the humidity channels are more significant than those for temperature. The calculated matrices contain contributions from the instrument error, the observation operator error and the RE. Due to the complex nature of observation error statistics, it is not known what portion of the error is RE. As humidity fields contain smaller-scale features than temperature fields, it is possible that it is the RE that contributes to the more significant error correlations.

In this article, we calculate horizontal REs using a method described by Daley (1993) and Liu and Rabier (2002). In Liu and Rabier (2002), the RE is calculated for a simple idealized system. Here we apply the Liu and Rabier (2002) method to real NWP data. We calculate the RE for temperature and humidity data over the UK. We consider the structure of RE to help understand whether significant correlations found in the work of Stewart (2010) and Stewart *et al.* (2009, 2012) can be attributed to RE. We investigate whether RE is more significant for humidity than temperature, and whether one approximation of RE error is suitable for all pressure levels.

In section 2 we describe the method used for calculating RE. We then describe the model and available data in section 3. Our experimental design is given in section 4 and we present our results in section 5. Finally we give conclusions in section 6.

## 2. Forward model error

### 2.1. Definition of forward model error

Forward model error,

$$\epsilon^H = \mathbf{y} - \mathcal{H}(\mathbf{x}^m), \quad (1)$$

is the difference between the noise-free observation vector,  $\mathbf{y}$ , of length  $p$  and the mapping of the exact model state vector,  $\mathbf{x}^m$ , of length  $N^m$  into observation space using the possibly nonlinear observation operator  $\mathcal{H}$ . The noise-free observation vector is a theoretical construct that represents an observation measured by a perfect observing instrument, i.e. with no instrument error. It is related to the actual measurement via the equation

$$\mathbf{y}^o = \mathbf{y} + \epsilon^I, \quad (2)$$

where  $\mathbf{y}^o$  is the observation vector and  $\epsilon^I$  is the unbiased instrument error.

The covariance of the forward model error  $E[(\epsilon^H - \bar{\epsilon}^H)(\epsilon^H - \bar{\epsilon}^H)^T] = \mathbf{R}^H$ , where the overbar  $\bar{\cdot}$  denotes the mean, is included in the observation-error covariance matrix  $\mathbf{R} = \mathbf{R}^H + \mathbf{R}^I$ , where  $\mathbf{R}^I = E[\epsilon^I \epsilon^{I^T}]$  is the instrument-error covariance matrix. As explained in section 1, forward model error is here defined as the sum of the error in the observation operator and the RE. In this study, because the observations are simulated, we are able to set the error in the observation operator to zero, and so the RE becomes the only source of forward model error. To calculate the REs in this article, we use a method defined by Daley (1993) and Liu and Rabier (2002). In this method it is assumed that the observations can be written as the mapping of a high-resolution state into observation space, and that the model state  $\mathbf{x}^m$  is a truncation of this high-resolution state.

### 2.2. Forward model error on a 1D domain

We restrict our calculations to the 1D periodic domain of length  $L = 2a\pi$ , where  $a$  is a constant which determines the length of the domain, and assume that the observation operator  $\mathbf{H}$  is linear. It is assumed that the high-resolution state  $x(r)$  at position  $r$  can be expressed as a Fourier series truncated at wave number  $K$ . At  $N$  points on the physical domain,  $-a\pi \leq r \leq a\pi$ , the function values  $x(r_j)$ ,  $j = 1, \dots, N$ , can be expressed in matrix form as

$$\mathbf{x} = \mathbf{F}\hat{\mathbf{x}}, \quad (3)$$

where  $\hat{\mathbf{x}}$  is a vector of length  $M = 2K + 1$  of spectral coefficients and  $\mathbf{F}$  is a Fourier transform matrix of dimension  $N \times M$ . In this work a number of Fourier matrices are used to calculate forward model error. A Fourier matrix  $\mathbf{F}$  of size  $m \times n$  has elements

$$F_{j,k} = \exp\left(\frac{2i k j \pi}{m}\right), \quad (4)$$

where  $j = 1, \dots, m$  and  $k = 1, \dots, n$ .

The model representation of the actual state is a wave-number-limited filter of the high-resolution state,  $\hat{\mathbf{x}}^m = \mathbf{T}\hat{\mathbf{x}}$  where  $\mathbf{T}$  is a truncation matrix which truncates the full spectral vector  $\hat{\mathbf{x}}$  to the analysed spectral vector  $\hat{\mathbf{x}}^m$ . The model representation of the actual state can be expressed as

$$\mathbf{x}^m = \mathbf{F}^m \hat{\mathbf{x}}^m, \quad (5)$$

where  $\hat{\mathbf{x}}^m$  is a vector of length  $M^m = 2K^m + 1$  of spectral coefficients and  $\mathbf{F}^m$  is a Fourier transform matrix of dimension  $N^m \times M^m$  with elements defined as in Eq. (4) but with no terms with wave number higher than  $K^m$ . The variables to be analysed are the Fourier spectral coefficients from  $-K^m$  to  $K^m$ ,  $K^m < K$ .

We define the observations by

$$y(r_o) = \int_{-a\pi}^{a\pi} x(r) w(r - r_o) dr. \quad (6)$$

Here the observations are defined as if they have been measured at point  $r_o$  by a remote-sensing instrument. The choice of the weighting function  $w(r)$  determines the type of observing

instrument. Writing Eq. (6) in spectral space allows us to write the  $p$  error-free observations as

$$\mathbf{y} = \mathbf{F}_p \mathbf{W} \hat{\mathbf{x}}, \quad (7)$$

where  $\mathbf{F}_p$  is a  $p \times M$  Fourier transform matrix and  $\mathbf{W}$  is a  $M \times M$  diagonal matrix with elements  $\hat{w}_k$ , the spectral coefficients of the weighting function  $w(r)$ .  $\mathbf{F}_p \mathbf{W}$  is an exact observation operator in spectral space. The measurement vector  $\mathbf{y}^o$  is given by

$$\mathbf{y}^o = \mathbf{F}_p \mathbf{W} \hat{\mathbf{x}} + \boldsymbol{\epsilon}^I. \quad (8)$$

The model representation of the observations is given by

$$\mathbf{y}^m = \mathbf{F}_p^m \mathbf{W}^m \mathbf{T} \hat{\mathbf{x}}, \quad (9)$$

where  $\mathbf{F}_p^m$  is the Fourier matrix with elements defined as in Eq. (4).  $\mathbf{W}^m$  is a  $M^m \times M^m$  diagonal matrix with elements  $\hat{w}_k^m$ , the spectral coefficients of the weighting function  $w(r)$ . This method assumes that the low-resolution model is a truncation of the high-resolution model. This allows forward model error to be considered in the perfect model case. It also allows us to exactly specify the observation operator so our forward model errors consist only of errors of representativity.

To obtain an equation for forward model error, we assume the observation operator is linear and substitute the definitions of observations, Eq. (7), and model representation of the observation, Eq. (9), into Eq. (1) to give

$$\boldsymbol{\epsilon}^H = \mathbf{F}_p \mathbf{W} \hat{\mathbf{x}} - \mathbf{F}_p^m \mathbf{W}^m \mathbf{T} \hat{\mathbf{x}}. \quad (10)$$

The expectation operation, denoted  $E[\dots]$ , is applied to give the forward model error covariance matrix,

$$\mathbf{R}^H = (\mathbf{F}_p \mathbf{W} - \mathbf{F}_p^m \mathbf{W}^m \mathbf{T}) \hat{\mathbf{S}} (\mathbf{F}_p \mathbf{W} - \mathbf{F}_p^m \mathbf{W}^m \mathbf{T})^*, \quad (11)$$

where  $\hat{\mathbf{S}} = E[(\hat{\mathbf{x}} - \bar{\mathbf{x}})(\hat{\mathbf{x}} - \bar{\mathbf{x}})^*]$  is the spectral covariance matrix for the high-resolution state and  $*$  denotes the complex conjugate transpose. The spectral covariance of the high-resolution state,  $\hat{\mathbf{S}}$ , contains information on how different wave numbers are related. It can be calculated using

$$\hat{\mathbf{S}} = \mathbf{F}^* \mathbf{S} \mathbf{F}, \quad (12)$$

where  $\mathbf{F}$  is a Fourier transform matrix and  $\mathbf{S} = E[(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^*]$  is the covariance matrix of the high-resolution state in physical space.

We now have an equation which can be used to calculate the RE covariance matrix for data on a periodic domain.

The method has been used previously to study the variance of forward model and RE for a simple static 1D system (Liu and Rabier, 2002) and there are limitations in using this method when applying it to NWP data. The method requires a periodic domain and assumes that the covariance of the state is isotropic and homogeneous, making it more applicable when calculating horizontal RE. The assumption that the truth is given by a high-resolution model is likely to cause an underestimate of RE as there will be scales which exist in the exact state, but which are not captured by the high-resolution model. Finally the method gives only a time-averaged estimate of forward model error.

Despite the limitations of this method, it is suitable to aid our understanding of the nature and structure of horizontal RE. Before using the method, we must define the weighting matrices to be used and describe how the spectral covariance of the high-resolution state can be calculated.

### 3. The model and data

In this study we calculate horizontal RE for both temperature and the log of specific humidity over the UK. The calculation of RE by the method of Liu and Rabier (2002) assumes that the actual state can be taken from a high-resolution model. As our actual state we take data from the Met Office UK variable resolution (UKV) model. The UKV model is a variable-resolution model that covers the UK. The model has a fixed regular grid on the interior with 1.5 km square grid boxes. The regular grid is surrounded by a variable-resolution grid where grid boxes smoothly increase in size to 4 km. For this study we consider two sets of data, previously used in Pavelin *et al.* (2009). The data cover sub-domains, each of 450 km  $\times$  450 km (300  $\times$  300 grid points with 1.5 km grid boxes), of the UKV model. The lateral boundary conditions for the 1.5 km models are taken from a 4 km grid-spacing regional model which is nested in the 12 km model which covers the North Atlantic and Europe (NAE). The boundary conditions blend into the 1.5 km model field over a transition zone of 10 km (Pavelin *et al.*, 2009) and we therefore exclude from our study the data in this region.

Since we are considering RE, it is also necessary to ensure that the model spectra have fully adjusted to the higher spatial resolution. This is not fully understood for this suite of models. However qualitative measures of the distance it takes convection to spin up due to features advecting in from the boundaries are given in Lean *et al.* (2008), Tang *et al.* (2012) and Kendon *et al.* (2012). We remove further data from the boundary so that approximately 30 km are removed in total. We expect the 1.5 km model to be spun up from the 4 km boundary conditions by this distance, although this would not be guaranteed for a rapidly changing synoptic situation.

In this work we calculate RE using the assumption that the model state is a truncation of high-resolution data. For the majority of our experiments, we chose a truncation factor which gives a model grid spacing equivalent to the grid spacing which is used in the Met Office NAE model. The Met Office NAE model has a grid spacing of 12 km (in midlatitudes) and covers Europe and the North Atlantic.

#### 3.1. The data available

We use temperature and humidity data over the UK available for two cases. The first case, Case 1, consists of data from 7 August 2007 at times 0830, 0900 and 0930 UTC on an area over the southern UK covering 3.04°W to 3.71°E and 49.18°N to 53.36°N. In this case there were partly clear skies with convection occurring over the southeast (Eden, 2007). The second set of data, Case 2, is from 6 September 2008 at 1400, 1430 and 1500 UTC covering 5.00°W to 1.20°E and 52.5°N to 56.00°N. In this case a deep depression tracked slowly eastnortheast across England (Eden, 2008). The data are available on a 300  $\times$  300 grid of latitude and longitude lines at each of 43 pressure levels.

#### 3.2. Creating samples from the data

There are some limitations to the data. Data near the boundary are contaminated by the boundary conditions taken from the coarser model. We remove this data at each pressure level by reducing the grid to a 256  $\times$  256 mesh centred on the main grid. We need to sample the data to calculate the covariance matrices for the actual state. The data we have are available on a 3D gridded domain. We are interested in calculating RE for individual pressure levels so for each experiment the data available are 2D; however, we are calculating RE on a 1D domain. To convert our data to 1D, we take the individual latitude rows of the data from the 749 hPa pressure level. We use this level as it is outside the boundary layer, but should still include the small-scale features which are relevant when calculating RE. We consider temperature and natural logarithm (ln) of specific humidity data for each of the



two synoptic cases. For each synoptic case, we have 256 samples at three different times and, therefore, we have 768 samples to calculate the covariance matrices. A covariance calculated with this number of samples is dominated by sampling error and hence this is not a sufficient number of samples to calculate an accurate representation of the required covariances. One way to overcome this would be to take data from more times. However, this would reduce the accuracy of the estimated RE at the specified time. A further problem is that the samples are not periodic, but the Liu and Rabier (2002) method assumes a periodic domain with a circulant covariance matrix  $\mathbf{S}$ . To overcome this and to increase the number of samples, we detrend and process the data.

### 3.3. Data processing

To create surrogate samples from each available sample, the data must be detrended. Detrending gives data on a homogeneous field; this is required by our chosen method for calculating RE. Data is detrended by removing a best-fit line using an appropriate polynomial of order no greater than 3 (Bendat and Piersol, 2011). It is justifiable to detrend the data as only trends with large length-scales are removed. All scales which contribute to the RE still remain. We detrend the 256 latitude samples at each available time. Different orders of polynomial were considered for detrending and the lowest-order polynomial which resulted in homogeneous data was chosen. A linear trend was removed from the temperature data, and a cubic trend from the natural logarithm of the specific humidity data. Removing polynomials of higher order had little effect on the RE results. These detrended data are now used to create new samples from each existing sample.

The method of Fourier randomization is used to generate surrogate samples from the same statistical distribution (Theiler *et al.*, 1992; Small and Tse, 2002). Fourier randomization consists of perturbing the phase of a set of data to create a new sample with a different phase, but where each wave number retains the same power. As the power spectrum of the sample is unchanged, the linear correlations are preserved. Therefore any choice of phase shift should result in data with the same covariance. As the covariance is preserved, we do not expect the choice of phase shift to affect the results when RE is calculated. Here we calculate circulant samples, which corresponds to shifting the phase of the data. This also gives the data the required periodicity. A circulant sample is created by shifting each element of the sample one position and taking the final element and making it the first entry in the sample. Each element can be shifted to each position, which means a sample with  $n$  elements can be used to create  $n$  circulant samples. Therefore creating surrogate samples increases the number of available samples we have for calculating the covariance of the high-resolution data. We have available 256 samples at three different times. Creating circulant samples gives us 65 536 samples at each time, and a total of 196 608 samples to estimate each of the covariance matrices, which is a sufficient number of samples.

## 4. Experiments

### 4.1. The covariance of the high-resolution data

We calculate the sample covariance matrix,  $\mathbf{S}$ , of the high-resolution data using

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^*, \quad (13)$$

where  $\mathbf{x}_i$  is the  $i$ th sample vector and  $\bar{\mathbf{x}}$  is a mean vector of the samples. We use this method and the model data to calculate the

Table 1. Variances for the true state at the 749 hPa pressure level.

	Temperature ( $\text{K}^2$ )	$\ln(\text{Specific humidity})$
Case 1	0.6638	0.0812
Case 2	0.1934	0.0178

The original unit for specific humidity is  $\text{kg kg}^{-1}$ .

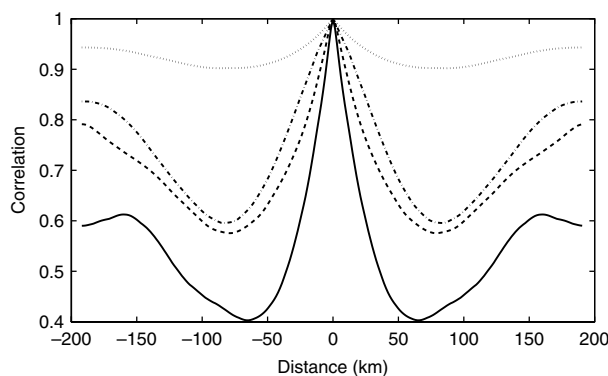


Figure 1. Correlation structure for the true fields of temperature and the natural logarithm of specific humidity at 749 hPa. Temperature: Case 1 dotted line, Case 2 dot-dash line. Specific humidity: Case 1 dashed line, Case 2 solid line.

covariance matrix of the actual state,  $\mathbf{S}$ . We use the circulant samples calculated from the UKV model data to calculate the covariance matrices for the temperature and natural logarithm of specific humidity fields for both cases. We give the variances in Table 1 and plot a row of each of the correlation matrices in Figure 1.

From Table 1 we see that the variances for Case 2 are smaller than those for Case 1. When considering the correlations plotted in Figure 1, we see that the temperature fields have larger correlations than the natural logarithm of specific humidity. For Case 1 the temperature correlations are very high; this is expected since, after detrending, this field is fairly constant. We also note that the correlations for Case 2 are smaller than the correlations for Case 1. This is due to the synoptic situation, since in Case 1 the field is more homogeneous with small-scale features over a small area of the domain, but in Case 2 the features have large-scale variations and are less homogeneous.

For the estimates of RE to be exact, we require the correct covariances of the truth. As our truth we are using data from the UKV model, and therefore our estimates of the covariances will only be as accurate as the spectra of the UKV model. As the UKV does not resolve all the scales in the truth, it is likely that the estimates of RE given by the Liu and Rabier (2002) method will be an underestimate. However as the UKV model is representative, in the characteristics of interest, of a reasonable truth and we are measuring the loss of information between the low- and high-resolution models, we can still expect to understand more about the behaviour and structure of RE.

### 4.2. The observations

To calculate the RE we require pseudo-observations. We expect RE to depend on observation type. To calculate these observation types we use Eq. (7), which requires a weighting matrix. We choose the weighting matrices in Eq. (11) to correspond to different types of observing instruments. The elements of the weighting matrix are the spectral coefficients of the weighting function  $w(r)$  which is used to define observations using Eq. (6). Pseudo-observations are created from the high-resolution data using three weighting functions. The weighting functions used here are the same as those used in Liu and Rabier (2002). Two of the weighting functions represent remotely sensed observations. One follows a top-hat (uniform) function with a width of 5 km. The other weighting function is calculated using a Gaussian curve with a width of

Table 2. Representativity error (RE) variances for temperature and natural logarithm of specific humidity at 749 hPa for Case 1. The values in brackets compare the RE variance to the high-resolution data variance.

Experiment number	Truncation	Number of observations ( $p$ )	Observation type	Temperature RE variance ( $K^2$ )	ln(Specific humidity) RE variance
1.1	32	32	Direct	$4.81 \times 10^{-3}$ (0.7%)	$1.51 \times 10^{-3}$ (1.9%)
1.2	32	32	Uniform	$2.71 \times 10^{-3}$ (0.4%)	$1.08 \times 10^{-3}$ (1.3%)
1.3	32	32	Gaussian	$8.99 \times 10^{-4}$ (0.1%)	$3.80 \times 10^{-4}$ (0.5%)
1.4	32	16	Direct	$4.81 \times 10^{-3}$ (0.7%)	$1.51 \times 10^{-3}$ (1.9%)
1.5	64	64	Direct	$2.13 \times 10^{-3}$ (0.3%)	$4.04 \times 10^{-4}$ (0.5%)

Table 3. As Table 2, but for Case 2.

Experiment number	Truncation	Number of Observations ( $p$ )	Observation type	Temperature RE variance ( $K^2$ )	ln(Specific humidity) RE variance
2.1	32	32	Direct	$2.21 \times 10^{-3}$ (1.1%)	$7.14 \times 10^{-4}$ (4.0%)
2.2	32	32	Uniform	$1.30 \times 10^{-3}$ (0.7%)	$4.84 \times 10^{-4}$ (2.7%)
2.3	32	32	Gaussian	$3.96 \times 10^{-4}$ (0.2%)	$1.60 \times 10^{-4}$ (0.9%)
2.4	32	16	Direct	$2.21 \times 10^{-3}$ (1.1%)	$7.14 \times 10^{-4}$ (4.0%)
2.5	64	64	Direct	$1.12 \times 10^{-3}$ (0.6%)	$2.50 \times 10^{-4}$ (1.4%)

20 km. We also consider *in situ* measurements. For these direct observations the weighting function  $w(r)$  in Eq. (6) becomes a Dirac delta-function. In this case the diagonal elements  $\hat{w}$  of the weighting matrix are all unity.

Now we have the appropriate weighting matrices and the covariance matrices for the high-resolution data at the 749 hPa pressure level. This allows us to calculate REs for temperature and natural logarithm of specific humidity. In the next section we present the results of our experiments.

## 5. Results

We now carry out a number of experiments to enable us to understand the nature of RE. The results for experiments carried out with data from Case 1 are given in Table 2, and for Case 2 in Table 3.

### 5.1. Temperature and humidity representativity errors

We first consider how the errors of representativity differ between the fields of temperature and of natural logarithm of humidity. We consider the RE for the case where the model has 32 points; this is a truncation of a factor of eight from the high-resolution model, which has 256 points. We start by assuming that we have direct observations. The values of the RE variance are given in Table 2, Experiment 1.1. We plot in Figures 2(a, b) (solid lines) the middle row of the RE correlation matrices for temperature and for natural logarithm of specific humidity from Case 1.

When we compare the variance of RE against the variance of the actual states, we see that RE is more significant for the natural logarithm of specific humidity than it is for temperature. We find that the RE variance for temperature is 0.7% of the high-resolution temperature variance, whereas the RE variance for natural logarithm of specific humidity is 1.9% of the high-resolution natural logarithm of specific humidity variance. When comparing the variance from this experiment with the same experiment carried out with Case 2 data (Table 3, Experiment 2.1), we see that the RE variances are smaller for Case 2. This is expected as there is less variance in the true fields in Case 2. However, these experiments show that the RE is more significant in this case, with RE for temperature being 1.1% of the high-resolution temperature variance and RE for natural logarithm of specific humidity being 4.0% of the high-resolution variance. For Case 1 from Figures 2(a, b) (solid lines) we see that the correlation structure is similar for both temperature and the natural logarithm of specific humidity. The correlations rapidly decrease in magnitude as the

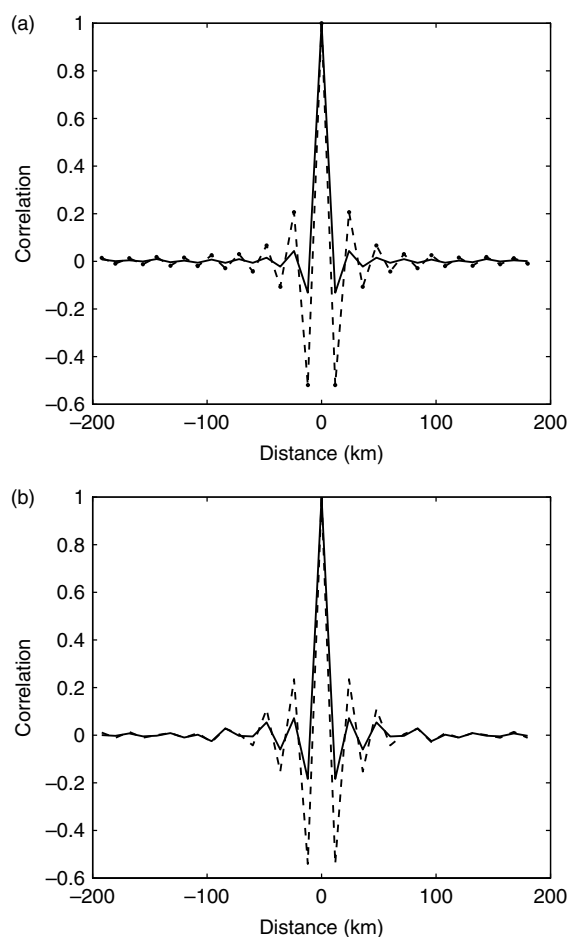


Figure 2. Representativity error correlations between observation centre points for Case 1 at 749 hPa with truncation to 32 points (12 km resolution) with every model grid point observed using direct (solid line) and Gaussian-weighted (dashed line) observations: (a) temperature (b) ln(specific humidity).

separation distance increases. The correlations for the natural logarithm of specific humidity are slightly larger, and decay less rapidly than the correlations for temperature.

### 5.2. Changing the observation type

We now consider what happens where the observations are defined with a uniform weighting matrix. This uniform

weighting acts on the temperature and natural logarithm of specific humidity fields. The variance of the RE is given in Table 2 Experiment 1.2. We see again, as expected, that the RE is more significant for the natural logarithm of specific humidity than it is for temperature. We see that the assumption of uniformly weighted observations has decreased the RE for both fields. This is as expected as the uniform observations do not capture all the small scales which the direct observations can. The correlations are larger than those when direct observations are used (result not shown). This is because two consecutive observations have some overlap in physical space. We see from Table 3 that Experiment 2.2 supports these results as the RE variance is smaller than that seen in Experiment 2.1.

We now consider what happens where the observations are defined using a Gaussian-weighting matrix. The results are given in Table 2 Experiment 1.3 and Table 3 Experiment 2.3. We plot the middle row of the RE correlation matrices for temperature and natural logarithm of specific humidity from Case 1 in Figures 2(a,b) (dashed lines). We find that the error variance is smaller than when either direct or uniform observations are assumed. Our Gaussian-weighted observations capture fewer small-scale features than the direct and uniform observations and hence the RE variance is smaller as the model captures a larger proportion of the scales captured by the observations. From the figures we see that the correlations for the RE calculated with these Gaussian-weighted observations are larger than the RE correlations present when direct observations are used. This is due to the overlapping of the weighting functions in physical space of nearby observations.

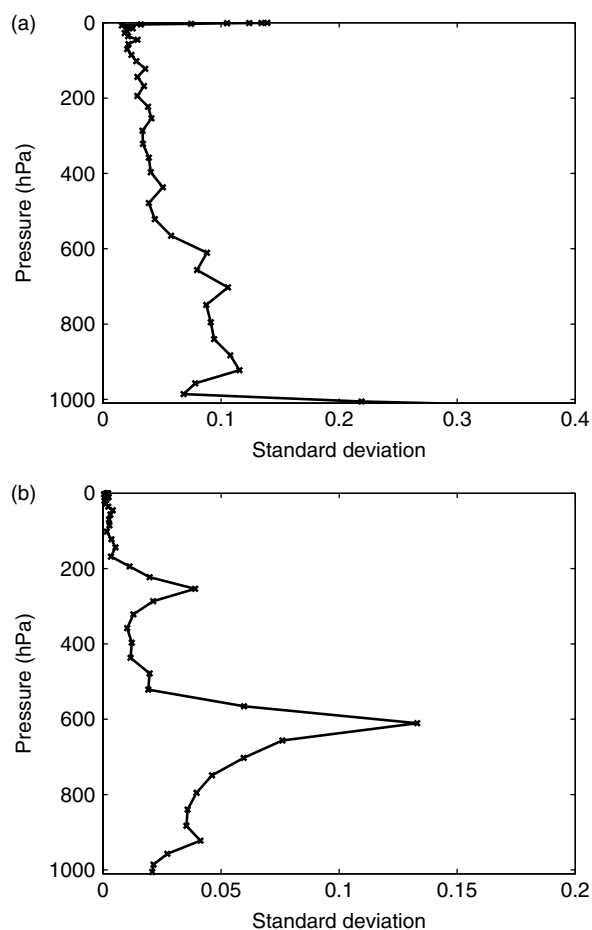
By comparing the experiments with different weighting functions, we see that the larger the weighting function length-scale used to define the observation, the lower the RE variance. Observations defined using weighting functions with larger length-scales capture fewer spatial scales. Therefore the difference between a larger length-scale observation and the model representation of the observation is smaller than a small length-scale observation and the model representation of the observation. Hence observations defined using weighting functions with larger length-scales result in smaller errors of representativity.

### 5.3. Number of observations

We now consider what happens when we calculate the RE where fewer direct observations are available. Experiments 1.4 in Table 2 and 2.4 in Table 3 show the error variance where only half the model grid points have associated direct observations. We see that having fewer observations available does not alter the variance of the RE. This is expected as RE applies individually to each observation and is independent of other observations. Experiments with uniform and Gaussian observations also support this conclusion. The Liu and Rabier (2002) method makes use of the Fourier transform; this leads to regularly spaced observations over the domain. The method used here leads to a class of correlation structures for RE which are dependent only on the distance between observations and not the number of observations available. We show this in the Appendix. Although the results in the Appendix are specific to the Liu and Rabier (2002) method, in general we expect that the RE variance should not be dependent on the number of available observations.

### 5.4. Number of model grid points

We now consider the results when the model has the larger number of 64 grid points. This is a higher truncation, so the model should be able to resolve more small-scale features, and hence we expect the errors of representativity to decrease. We give the results for experiments with direct observations in Table 2 Experiment 1.5 and Table 3 Experiment 2.5. From these results we see that the RE variances have decreased. For direct observations



**Figure 3.** Representativity error standard deviation with model level height, for Case 1 with 32 direct observations (every model grid point observed): (a) temperature, (b)  $\ln(\text{specific humidity})$ .

the RE has been approximately halved. Experiments with uniform and Gaussian weightings are not shown, but produce results that also support this conclusion.

### 5.5. Representativity errors at different model levels

So far we have considered the RE only at the 749 hPa model level height. We now calculate a RE for each pressure level of the model. This will allow us to consider the variation of RE with height. From this we can determine if one realisation of RE would be suitable at every pressure level, or if it is more appropriate to use the correct RE for each level.

Before calculating the RE for each model level, we must first calculate the covariance matrices for the high-resolution data for temperature and natural logarithm of specific humidity for each pressure level. We use the same data, but at the correct pressure level, and the same preprocessing techniques described in section 3.

We consider the case where we have truncated to 32 grid points and have 32 direct observations available. We plot the standard deviation of RE for Case 1 in Figures 3(a) (temperature) and 3(b) (natural logarithm of specific humidity) and for Case 2 in Figures 4(a) (temperature) and 4(b) (natural logarithm of specific humidity).

From the figures we see that RE for temperature is more constant with height than that for the natural logarithm of specific humidity. The exception to this is in the boundary layer, where the temperature RE is large. For the natural logarithm of specific humidity in Case 1, we see a large increase in the RE standard deviation between 749 and 610 hPa. For Case 2 the largest peak in RE is seen at 300 hPa. These levels are where cloud is seen and hence it is at these levels where the small-scale humidity features



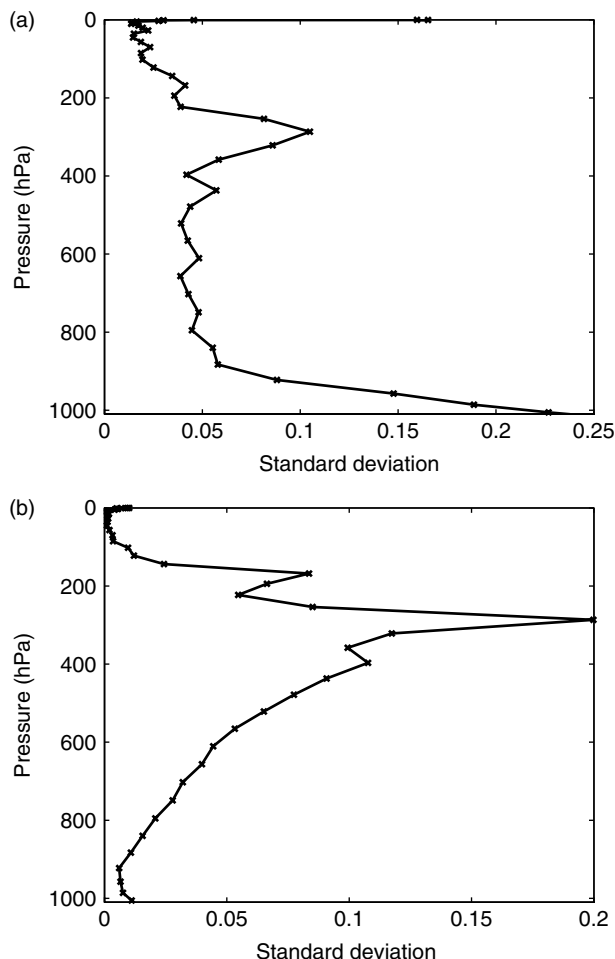


Figure 4. As Figure 3, but for Case 2.

exist; this results in the larger RE variances. Finally we consider how the correlation structure varies with height. We found that for both temperature and specific humidity at different pressure levels the correlation structures of the REs were qualitatively similar to those for the 749 hPa level, as seen in Figure 2. The difference in variance and minimal difference in correlation structure can be attributed to the different scales in the true state which are represented in  $\hat{\mathbf{S}}$ , used to calculate the RE (Appendix, Eq. (A.4)).

The results show that RE is not constant with height and in some cases it may be beneficial to have a RE matrix where the variance varies with height. They also support our conclusions that RE is strongly case-dependent. Experiments (results not shown) with uniform- and Gaussian-weighting functions and truncation to 64 points also show that RE varies with height.

Here we have shown that the horizontal RE varies with height. We have not considered whether REs are vertically correlated as it is beyond the scope of this article. Also is not obvious how to apply the Liu and Rabier (2002) method to vertical data since the model levels are not equally spaced. Although vertical REs are not calculated here, we would be surprised if RE is not correlated in the vertical.

## 6. Conclusions

In this study we use a method defined in Daley (1993) and Liu and Rabier (2002) to calculate RE. Previously the method has been used to investigate RE for a simple system. We adopt a new approach by applying the method to NWP data. We wished to investigate whether significant correlations in the observation-error matrix (Stewart *et al.*, 2009, 2012; Stewart, 2010) could be attributed to RE, and whether RE is more significant for

the natural logarithm of specific humidity than for temperature. We calculated and compared REs for temperature and natural logarithm of specific humidity. To calculate the RE it is necessary to have an estimate of the covariance of the truth state. Here this covariance is calculated using data from the Met Office UKV model. The accuracy of the RE estimates depend on the accuracy of these covariances and, as the UKV model cannot represent all the scales in the truth, it is possible that the RE is underestimated. Experiments using data from the Met Office UKV model showed that RE was more significant for the natural logarithm of specific humidity than for temperature. This was determined by comparing the size of the RE variance with the variances of the high-resolution states. This suggests that correlations found in previous approximations of the observation-error covariance matrix,  $\mathbf{R}$ , such as those in Stewart (2010) and Weston (2011), are likely to be RE, at least in part. We calculated RE using data from two different cases and showed that real data RE is sensitive to the synoptic situation, which supports claims by Janjic and Cohn (2006). We also found that, as the number of model grid points is reduced, the RE increases. This is because, at lower resolution, the model is not able to resolve as many scales. We also found that using direct observations gave a higher RE than when uniform- or Gaussian-weighted observations were used. This is because the direct observations contain more information on smaller scales than the uniform- or Gaussian-weighted observations. Experiments showed that altering the number of observations used to calculate RE had no effect on the RE variance. We showed that this method leads to a class of correlation structures which depends only on the distance between observations and not the number of observations. We believe that in general the number of observations should not affect the RE correlation structure and that the structure is dependent only on the distance between the available observations. Finally we considered how the RE standard deviation varied at different pressure levels. We found that representativity does vary at different pressure levels and this means that assumptions such as those in Dee and Da Silva (1999), where errors at different model levels are fixed, may not be suitable when RE is taken into account. As it becomes more efficient to use correlated observation errors in data assimilation systems, good approximations of the observation-error covariance matrix will be required. Using the method of Liu and Rabier (2002), we have calculated REs for specific fields and have shown that errors of representativity are correlated. However further work is required to determine whether the inclusion of these errors in an assimilation scheme improves the analysis.

## Acknowledgements

The authors thank Laura Stewart for providing the data. This work was funded by NERC as part of the National Centre for Earth Observation and by the Met Office through a CASE studentship.

## Appendix

### Representativity error variance

Here we show why the RE variance does not alter when calculated with different numbers of observations. We do this by considering the calculation of the elements of the RE matrix.

RE is calculated using Eq. (11). The matrices are:

- $\mathbf{F}$  with  $j = 1 \dots p$  and  $k = 1 \dots M$ .  
Elements defined as in Eq. (4).
- $\mathbf{F}^m$  with  $j = 1 \dots p$  and  $k = 1 \dots M^m$ .  
Elements defined as in Eq. (4).
- $\mathbf{W}$  with  $j = 1 \dots M$  and  $k = 1 \dots M$ .  
Elements  $\hat{w}_j$  when  $j = k$ , 0 otherwise.
- $\mathbf{W}^m$  with  $j = 1 \dots M^m$  and  $k = 1 \dots M^m$ .  
Elements  $\hat{w}_j$  when  $j = k$ , 0 otherwise.

- $\widehat{\mathbf{S}}$  with  $j = 1 \dots M$  and  $k = 1 \dots M$ .  
Elements  $\widehat{s}_j$  when  $j = k$ , 0 otherwise.
- $\mathbf{T}$  with  $j = 1 \dots M^m$  and  $k = 1 \dots M$ .  
Elements 1 when  $j = k$  and  $k \leq M$ , 0 otherwise.

First we calculate the elements of  $\mathbf{A} = \mathbf{FW}$  and  $\mathbf{B} = \mathbf{FWT}$ . As many of the elements are zero, we find that  $A_{j,k} = F_{j,k}W_{jj}$  and  $B_{j,k} = F_{j,k}W_{jj}^m$  when  $k \leq M^m$ , and 0 otherwise.

Next we calculate elements of  $\mathbf{C} = \mathbf{A} - \mathbf{B}$ ,

$$C_{j,k} = \begin{cases} \exp\left(\frac{2ikj\pi}{p}\right)\widehat{w}_j & M^m < k \leq M, \\ 0 & \text{otherwise.} \end{cases} \quad (\text{A.1})$$

Now we calculate  $\mathbf{R}^H = \widehat{\mathbf{C}}\widehat{\mathbf{S}}^*$ . Elements of  $\mathbf{E} = \widehat{\mathbf{C}}\widehat{\mathbf{S}}$  are  $E_{j,k} = C_{j,k}\widehat{s}_{k,k}$ . Finally we calculate  $\mathbf{R}^H = \mathbf{E}\mathbf{C}^*$ , where  $\cdot^*$  is the complex conjugate transpose and  $\cdot^{\cdot}$  is the complex conjugate,

$$R_{j,k}^H = \sum_{l=0}^M E_{j,l}C_{l,k}^*, \quad (\text{A.2})$$

$$= \sum_{l=1}^M C_{j,l}\widehat{s}_{l,l}\overline{C}_{k,l}, \quad (\text{A.3})$$

$$= \sum_{l=1}^M \exp\left(\frac{2ijl\pi}{p}\right)\widehat{w}_l\widehat{s}_l\overline{\widehat{w}_l} \exp\left(\frac{-2ikl\pi}{p}\right). \quad (\text{A.4})$$

We now show that the variance does not change when  $p$  changes. This is the case when  $j = k$ ; that is,

$$R_{j,j}^H = \sum_{l=1}^M \exp\left(\frac{2ijl\pi}{p}\right)\widehat{w}_l\widehat{s}_l\overline{\widehat{w}_l} \exp\left(\frac{-2ijl\pi}{p}\right), \quad (\text{A.5})$$

$$= \sum_{l=1}^M \widehat{w}_l\widehat{s}_l\overline{\widehat{w}_l}. \quad (\text{A.6})$$

This does not depend on  $p$  and hence we do not expect the variance to change when we use different numbers of observations to calculate RE.

We now show that the correlation structure depends only on the distance between observations and not the number of observations.

Our model has  $N^m$  grid points separated by a spacing  $\Delta x$  and we have  $p$  observations. The distance between consecutive observations is  $N^m \Delta x / p$ . Suppose we have two observations separated by a distance  $d$  and assume that these are observation  $j$  and observation  $k$ . Then we have

$$d = \frac{(j-k)(N^m \Delta x)}{p}, \quad (\text{A.7})$$

and hence

$$\frac{j-k}{p} = \frac{d}{N^m \Delta x}. \quad (\text{A.8})$$

Substituting this into Eq. (A.4), we obtain

$$R_{j,k}^H = \sum_{l=1}^M \widehat{w}_l\widehat{s}_l\overline{\widehat{w}_l} \exp\left\{\frac{2il\pi(j-k)}{p}\right\}, \quad (\text{A.9})$$

$$= \sum_{l=1}^M \widehat{w}_l\widehat{s}_l\overline{\widehat{w}_l} \exp\left\{\frac{2il\pi d}{N^m \Delta x}\right\}. \quad (\text{A.10})$$

Hence the correlations depend only on the distance between the observations and not on the number of observations.

## References

- Bendat JS, Piersol AG. 2011. *Random Data: Analysis and Measurement Procedures*. John Wiley & Sons: Chichester, UK.
- Bormann N, Bauer P. 2010. Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. I: Methods and application to ATOVS data. *Q. J. R. Meteorol. Soc.* **136**: 1036–1050.
- Bormann N, Saariene S, Kelly G, Thépaut J-N. 2002. The spatial structure of observation errors in atmospheric motion vectors from geostationary satellite data. *Mon. Weather Rev.* **131**: 706–718.
- Bormann N, Collard A, Bauer P. 2010. Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. II: Application to AIRS and IASI data. *Q. J. R. Meteorol. Soc.* **136**: 1051–1063.
- Cohn SE. 1997. Introduction to estimation theory. *J. Meteorol. Soc. Japan* **75**: 275–288.
- Daley R. 1991. *Atmospheric Data Analysis*. Cambridge University Press: Cambridge, UK.
- Daley R. 1993. Estimating observation error statistics for atmospheric data assimilation. *Ann. Geophys.* **11**: 634–647.
- Dee DP, Da Silva AM. 1999. Maximum-likelihood estimation of forecast and observation error covariance parameters. Part I: Methodology. *Mon. Weather Rev.* **127**: 1822–1843.
- Desroziers G, Berre L, Chapnik B, Poli P. 2005. Diagnosis of observation, background and analysis-error statistics in observation space. *Q. J. R. Meteorol. Soc.* **131**: 3385–3396.
- Eden P. 2007. Weather Log: August 2007. *Weather* **62**: i–iv.
- Eden P. 2008. Weather Log: September 2008. *Weather* **63**: i–iv.
- Fisher M. 2005. ‘Accounting for correlated observation error in the ECMWF analysis’. Tech. report MF/05106. ECMWF: Reading, UK.
- Harris BA, Kelly G. 2001. A satellite radiance-bias correction scheme for data assimilation. *Q. J. R. Meteorol. Soc.* **127**: 1453–1468.
- Healy SB, White AA. 2005. Use of discrete Fourier transforms in the 1D-Var retrieval problem. *Q. J. R. Meteorol. Soc.* **131**: 63–72.
- Hilton F, Collard A, Guidard V, Randriamampianina R, Schwaerz M. 2009. ‘Assimilation of IASI Radiances at European NWP Centres’. In *Proceedings of workshop on the assimilation of IASI data in NWP*, 6–8 May 2009. 39–48. ECMWF: Reading, UK.
- Hollingsworth A, Lönnerberg P. 1986. The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: The wind field. *Tellus A* **38**: 111–136.
- Houtekamer PL, Mitchell HL. 2005. Ensemble Kalman filtering. *Q. J. R. Meteorol. Soc.* **133**: 3260–3289.
- Janjic T, Cohn SE. 2006. Treatment of observation error due to unresolved scales in atmospheric data assimilation. *Mon. Weather Rev.* **134**: 2900–2915.
- Kendon EJ, Roberts NM, Senior CA, Roberts MJ. 2012. Realism of rainfall in a very high-resolution regional climate model, *J. Climate* **25**: 5791–5806.
- Lahoz WA, Khattatov B, Menard R. (eds) 2010. *Data Assimilation; Making Sense of Observations*. Chapter 2. Springer: Berlin.
- Lean HW, Clark PA, Dixon M, Roberts NM, Fitch A, Forbes R, Halliwell C. 2008. Characteristic of high-resolution versions of the Met Office Unified Model for forecasting convection over the United Kingdom. *Mon. Weather Rev.* **136**: 3408–3424.
- Liu Z-Q, Rabier F. 2002. The interaction between model resolution observation resolution and observation density in data assimilation: A one-dimensional study. *Q. J. R. Meteorol. Soc.* **128**: 1367–1386.
- Lorenc AC. 1986. Analysis methods for numerical weather prediction. *Q. J. R. Meteorol. Soc.* **112**: 1177–1194.
- Pavelin E, English S, Bornemann FJ. 2009. ‘MTG-IRS Simulation Project’. EUMETSAT study contract EUM/CO/08/4600000492/SAT. Met Office: Exeter, UK. [http://www.eumetsat.int/groups/pps/documents/document/pdf-mtg\\_rep36.pdf](http://www.eumetsat.int/groups/pps/documents/document/pdf-mtg_rep36.pdf).
- Sherlock V, Collard A, Hannon S, Saunders R. 2003. The Gastropod fast radiative transfer model for advanced infrared sounders and characterization of its errors for radiance assimilation. *J. Appl. Meteorol.* **42**: 1731–1747.
- Small M, Tse CK. 2002. Applying the method of surrogate data to cyclic time series. *Physica D* **164**: 187–201.
- Stewart LM. 2010. ‘Correlated observation errors in data assimilation’. PhD thesis, Department of Mathematics and Statistics, University of Reading, UK.
- Stewart LM, Dance SL, Nichols NK. 2008. Correlated observation errors in data assimilation. *Int. J. Numer. Meth. Fluids* **56**: 1521–1527.
- Stewart LM, Cameron J, Dance SL, English S, Eyre JR, Nichols NK. 2009. ‘Observation error correlations in IASI radiance data’. Mathematics reports series 1/2009, University of Reading: UK. [http://www.reading.ac.uk/web/FILES/maths/obs\\_error\\_IASI\\_radiance.pdf](http://www.reading.ac.uk/web/FILES/maths/obs_error_IASI_radiance.pdf).
- Stewart LM, Dance SL, Nichols NK, Eyre JR, Cameron J. 2012. ‘Estimating interchannel observation error correlations for IASI radiance data in the Met Office system’. Preprint MPS-2012-26, Department of Mathematics

- and Statistics, University of Reading: UK. <http://www.reading.ac.uk/maths-and-stats/research/maths-preprints.aspx>.
- Stewart LM, Dance SL, Nichols NK. 2013. Data assimilation with correlated observation errors: experiments with a 1-D shallow-water model. *Tellus A* **65**: 19546.
- Tang Y, Lean HW, Bornemann J. 2012. The benefits of the Met Office variable resolution NWP model for forecasting convection. *Meteorol. Appl.* DOI: 10.1002/met.1300. In press.
- Theiler J, Eubank S, Longtin A, Galdrikian B, Farmer JD. 1992. Testing for non-linearity in time series: the method of surrogate data. *Physica D* **58**: 77–94.
- Weston P. 2011. 'Progress towards the Implementation of Correlated Observation Errors in 4D-Var'. Forecasting Research Technical Report 560. Met Office: Exeter, UK.
- Whitaker JS, Hamill TM, Wei X, Song Y, Toth Z. 2008. Ensemble data assimilation with the NCEP Global Forecast System. *Mon. Weather Rev.* **136**: 463–482.